Using Mutual Information to Measure the Impact of Multiple Genetic Factors on AIDS

George W. Nelson, PhD, and Stephen J. O'Brien, PhD†*

Summary: Since the discovery of the 32–base-pair deletion in the *CCR5* chemokine receptor gene (*CCR5-Δ32*) and its effect on HIV-1 infection and AIDS progression, many genetic factors affecting AIDS have been identified. Here we quantify the impact of 13 of these factors on AIDS progression using a new statistic based on the mutual information between causal factors and disease, the explained fraction. The influence of causal factors on disease is commonly measured by the attributable fraction statistic, but the attributable fraction is a poor measure of the extent to which a factor explains disease because it considers only whether a factor is necessary, not whether it is sufficient. The definition of the explained fraction, which is analogous to R^2 or the explained variation for regression models, extends naturally to multiple factor levels. Because the explained fraction is approximately additive, it can be used to estimate how much of epidemiological data is explained by known genetic or environmental factors, and conversely how much is yet to be explained by unknown factors. We show that 13 genetic factors can cumulatively explain 9% of slow progression to AIDS, an effect comparable to the effect of smoking on lung cancer.

Key Words: attributable fraction, mutual information, explained variation, multifactorial disease, AIDS host genetic factors

(*J Acquir Immune Defic Syndr* 2006;42:347–354)

The importance of a factor causing or contributing to disease is a function both of the frequency of the factor and of the strength of its effect. A standard measure of the impact of a factor is the attributable fraction (AF) (alternatively, attributable risk or population attributable risk) defined as the fraction of individuals in the population with a given outcome (eg, a specified disease) whose condition can be attributed to a given risk factor.[1,2] In terms of the relative risk (R) and the frequency (f) of the factor, $AF = f(R-1)/(1+f(R-1))$.

The definition of the AF arises naturally from the epidemiological search for exposures causing disease, for example, smoking as a factor causing lung cancer or heart disease. The AF is a particularly useful measure for potentially controllable factors because it gives the fraction of the disease that would be prevented if the factor were eliminated. However, because the AF measures only whether a factor is necessary for disease, not whether it is sufficient, it is a poor measure of the extent to which known factors explain disease for several reasons. First, the AF is inherently asymmetric: it measures how much of one disease state can be attributed to a given factor, but not how much of the absence of the disease state can be attributed to the absence of the factor. For the case of smoking and disease, it is clear that the smoking is the exposure, but for genetic polymorphisms affecting disease, it is not in general clear which allele or genotype constitutes the exposure. Second, the AF is in general not at all additive: if 2 factors are each necessary for disease, each will have an AF of 100%. However, it would be useful for a scientific measure of understanding to be approximately additive, running from 0% (no understanding) to 100% (complete understanding or ability to predict).

The influence of *CCR5-Δ32* on HIV-1 infection provides a good example of the asymmetry of the AF. Homozygous *CCR5-Δ32* people are virtually immune to infection because they lack the requisite cellular receptor, CCR5, for cell entry by the primary infecting strains of HIV-1. Because this genotype is rare, the AF for *CCR5-Δ32* homozygosity for protection from infection after multiple exposures is only 5%.[3,4] However, because HIV infection is in fact the disease state, it is reasonable to pose the question in the other way: how much of HIV infection can be attributed to the presence of a functional CCR5 receptor (ie, to the absence of the Δ32 mutation from at least one chromosome); defined this way, we obtain an AF of 99%. However, if we are concerned with measuring understanding, explaining susceptibility or protection from infection should give the same answer.

Third, the standard AF calculation requires dividing the individuals into dichotomous groups, exposed versus unexposed subjects. With multiple factors, an AF may be defined by comparing all subjects with one or more susceptible factors with subjects with none;[5] with protective factors, we may calculate an AF for protection in the same way. However, if there are multiple factors independently affecting disease, there will be multiple levels of protection or susceptibility; splitting the subjects into 2 groups does not account for the varying effects of the factors. Moreover, this approach cannot deal with the case of both protective and

Received for publication January 11, 2005; accepted February 20, 2006.
From the *Basic Research Program, Science Applications International Corporation Frederick, National Cancer Institute (NCI) Frederick, MD; and †Laboratory of Genomic Diversity, NCI, Frederick, MD.
Sponsorship: funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under contract No. NO1-C0-12400.
Reprints: George W. Nelson, PhD, Basic Research Program, SAIC Frederick, Inc., Bldg. 560, Room 21-64, Frederick, MD 21702-1201 (e-mail: nelsong@ncifcrf.gov).

susceptible factors. We might compare all other subjects with the most protected group, but if there is a rare, highly protected group, we get the same implausible result as the calculation of AF for not being *CCR5-Δ32* homozygous. Extending the concept of the AF to multiple levels in a rigorous way is possible,[6] but necessarily involves a rather complex algorithm.

An alternative is to use a measure of correlation of disease and factors. For ordinary linear regression, the coefficient of determination R^2 measures the ability of the factors to predict outcomes. Schemper has extensively developed a statistic called the explained variation (EV) which generalizes R^2 and can be calculated for more general regression models, in particular for logistic regression and the Cox model.[7–9]

The mutual information is a basic measure of prediction or correlation between factors and outcomes.[10,11] Here we present a new statistic, the explained fraction (EF), which uses mutual information to quantify the effect of causal factors on disease. The EF is analogous to R^2 or the EV, but is specifically defined for a contingency table and thus allows direct comparison to the AF. We apply the EF to a number of epidemiological examples, in particular to the multifactorial case of the genetic epidemiology of AIDS progression.

DEFINITION OF THE EF

Mutual Information

Whereas the AF is a measure of the importance of a factor for occurrence of disease, an alternate approach is to consider the informativeness of knowledge of the factor in predicting the occurrence of disease. The mutual information between causal factors and disease states is a specific measure of this. Here we consider only cases where both the causal factors and the disease outcomes are described by categorical variables, so that their relationship is described by a contingency table $[a_{ij}]$ of frequencies, with the rows i representing causal factors and the columns j representing disease categories. For multiple causal factors, each combination of factors is represented by a row. We assume initially that there are 2 disease categories, $j = 1, 2$, but there are an arbitrary number of factors or combinations of factors, $i = 1$ to n. The row marginals $\sum_{j=1}^{2} a_{ij} = a_i$ correspond to the frequency of the factor state i, whereas the column marginals $\sum_{i1}^{n} a_{ij} = a_j$ correspond to the frequency of the disease category j.

The mutual information between the rows and the columns of a contingency table is given by:[10]

$$I(1,2) = \sum_{i,j} a_{ij} \log \frac{a_{ij}}{a_{i\cdot} a_{\cdot j}}. \quad (1)$$

The notation $I(1,2)$ indicates that this is the information obtained from sampling a single individual from the population, discriminating between the hypotheses that the rows and columns are dependent (1) or independent (2). For our case where the rows represent the causal factors and the columns represent the disease categories, $I(1,2)$ is a measure of the information about the probabilities of disease outcomes that we gain by knowing the factors. Recalling that the well-known likelihood ratio statistic L is given by

$$L = \sum_{i,j} A_{ij} \log \frac{A_{ij}}{N a_{i\cdot} a_{\cdot j}} = N \sum_{i,j} a_{ij} \log \frac{a_{ij}}{a_{i\cdot} a_{\cdot j}}, \quad (2)$$

where A_{ij} are the number of subjects observed for the i, j cell of the contingency table, we note that $I(1,2) = L/N$.

Explained Fraction

Consider a case in which knowledge of the causal factors completely predicts an individual's disease status, for example, the case of an inherited disease caused by a single genotype with complete penetrance. $I(1,2)$ in this case has a finite value, calculated below, which is a function only of the frequencies of the disease categories. Because this represents complete knowledge, it is reasonable to normalize the mutual information measure by dividing $I(1,2)$ by this maximal value to give a measure that ranges from 0 (corresponding to no information obtained) to 1 (corresponding to a total ability to predict). Thus, this is a measure of the fractional extent to which the disease is explained by knowledge of the causal factors, which we call the EF.

To calculate the denominator of the EF, we observe that if knowledge of causal factors completely predicts disease outcomes, then each factor status must be totally associated with a single disease status, so each row i of the table has a single nonzero entry which must equal the row marginal $a_{i\cdot}$. Taking $0*\log(0) \equiv 0$, Eq. (1) becomes:

$$\begin{aligned} I_{\max}(1,2) &= \sum_i a_{i1} \log\left(\frac{a_{i\cdot}}{a_{i\cdot} a_{\cdot 1}}\right) + \sum_i a_{i2} \log\left(\frac{a_{i\cdot}}{a_{i\cdot} a_{\cdot 2}}\right) \\ &= -a_{\cdot 1} \log(a_{\cdot 1}) - a_{\cdot 2} \log(a_{\cdot 2}) \end{aligned} \quad (3)$$

Because the column marginals are simply the frequencies of the disease states D_1, D_2, we have $I_{\max}(1,2) = -D_1 \log(D_1) - D_2 \log(D_2)$, which is the average information required to specify the disease state of an individual, or the entropy of the disease states. More generally, we can have n disease states, so $I_{\max}(1,2) = -\sum_j D_j \log(D_j)$. Thus, the EF is given by

$$EF = \frac{\sum_{i,j} a_{ij} \log\left(\frac{a_{ij}}{a_{i\cdot} a_{\cdot j}}\right)}{-\sum_j D_j \log(D_j)}. \quad (4)$$

We consider in Appendix 1 the relationship of the EF to the regression R^2, and in Appendix 2 some statistical properties of the EF. Appendix 2 presents an approach to calculating an unbiased estimator and confidence intervals for the EF. We show in Appendix 1 that the EF is in fact analogous to R^2 and thus may be considered an EV calculated for a contingency table. We use the name "explained fraction" (EF) because this emphasizes the analogy with the AF. In general, logistic regression is an alternative, often advantageous, to a contingency table calculation; we consider the contingency table calculation here because it allows a direct comparison of the AF. However, for any population study (as opposed to a case control study), the EV for the disease phenomena may reasonable be considered to be a measure of the EF.

EF EXAMPLES

To illustrate the calculation and the interpretation of the EF, we consider 2 non-AIDS diseases whose epidemiology is notably well characterized.

Tangier Disease

Tangier disease is an autosomal recessive disease characterized by the absence of high-density lipoprotein in plasma, with the most visible symptom being extremely enlarged tonsils due to excessive lipid deposits.[12] The ATP-binding cassette transporter gene *ABC1* was identified by chromosomal mapping and biochemical function as a candidate gene for Tangier disease.[13,14] Examination of 5 Tangier disease pedigrees showed that 10 of 10 individuals with Tangier disease had mutated *ABC1* genes on both chromosomes, whereas 26 of 26 individuals in the same pedigrees lacking disease had normal *ABC1* genes on one or both chromosomes.[15] Thus, for the available data, the disease is completely explained. Formally, we have the contingency table $\begin{bmatrix} 26 & 0 \\ 0 & 10 \end{bmatrix}$; there is only one nonzero entry per row, so EF = 1. The AF is also 1, so in this case, EF = AF. The ability of a single factor to explain a disease completely stems from the influence of causative functional mutational variants in a critical physiological process. Other monogenic diseases exist, but complex diseases generally have multiple interaction genetic and environmental causes.

Cigarette Smoking and Lung Cancer

We consider the 33-year follow-up on the Swedish Smoking Habit Survey, which represents an epidemiological assessment of 12,664 Swedish men at risk for lung cancer.[16] Lung cancer deaths were considered together with those from cancer of the trachea and bronchus; among males, there were 36 deaths among 8156 respondents who had never smoked, and 177 deaths among 4508 current smokers. We first calculate the AF of smoking for these cancers; using the standard definition, we are forced to decide whether to consider former smokers as smokers or nonsmokers (neither of which is really satisfactory) or to exclude them from the analysis. Excluding them from the analysis, the frequency of the exposure (current smoking) is 0.356, the published relative risk for this study (from a stratified analysis) is 9.40, so AF = 0.745; that is, 75% of lung and respiratory tract cancer deaths can be attributed to smoking. For the same contingency table, by Eq. (4), EF = 0.096, indicating that 9.6% of the occurrence or nonoccurrence of lung and respiratory tract cancer can be explained by smoking (as compared with never having smoked). Because the EF is defined for multiple exposures, we can also consider different levels of smoking and can include former smokers as a separate category. With this additional information, EF = 11.5%.

Both of these analyses give an EF much smaller than the AF for being a current smoker. The AF is large because it reflects only the fact that almost all of those who die of lung cancer are smokers. The EF also reflects the fact that most of those who smoke never get lung cancer (it is estimated that even those who smoke 30 cigarettes per day for 50 years have only a 1 in 6 chances of dying from lung cancer[17]). Clearly, there are additional causal influences yet to be discovered, perhaps including genetic susceptibility factors, that determine susceptibility to lung cancer among smokers.

EF FOR GENETIC FACTORS AFFECTING AIDS PROGRESSION

AIDS Long-Term Survival Versus Progression

The influence of host genetic factors (AIDS restriction genes, ARGs) on AIDS progression is an important case of multiple factors. We consider the effects of 13 genetic factors that have been typed on AIDS cohorts as influences on slow progression to AIDS.[18–21] We define the longest surviving third of subjects, for each of 4 AIDS definitions, to be slow progressors. Of the 13 factors, 11 have 2 levels (in most cases for dominant or recessive influence on AIDS progression) and 2 have 3 or more levels, giving 16,000 possible genotypic combinations, of which 162 actually occur in the study group, clearly too many for robust results.

The 13 genetic factors have an independent effect on AIDS progression in that their effects on survival are additive, where factors are significantly nonadditive, for example, for the *KIR 3DS1-HLA Bw4 80I* interaction, a term for the interaction is included in the model. Therefore, rather than considering the empirical matrix whose entries are the actual numbers of slow progressor subjects with each combination of factors, we consider a smoothed distribution given by 2 approximations: first, that progression to the AIDS endpoint is described by a Weibull distribution; second, that the survival effect of the factors is additive (although the Weibull distribution may not precisely describe AIDS progression, the fit is extremely good, and the Weibull allows considering factors as additive). With these assumptions, the survival distribution of each group of factors is given by the Weibull distribution with the same shape parameter as for the whole group and the sum of the β value for the factors as scale parameter. This survival distribution determines the fraction of subjects with the given combination of factors that progress before the cutoff time for slow progression versus those who survive past this time, that is, the relative frequencies for each row of the contingency table.

Table 1 shows the calculated EF for long-term AIDS versus more rapid progression, for the 13 individual genetic factors, and for combined factors, along with AFs calculated with the same approximations. The EF values for given factors vary greatly between outcomes, illustrating the fact the some factors have the greatest effect on early outcomes and others on late outcomes. Focusing on progression to clinical AIDS symptoms (the 1987 CDC AIDS definition), the EFs for individual factors range from 0.1% to 2.3% for the protective factor HLA B*27. The overall EF is calculated for the 162 by 2 table in which each combination of factors is represented by a row. For progression to AIDS 1987, the EF for combined factors is 9.0%, somewhat less than the sum of the EFs for individual factors (9.4%). The difference is small, demonstrating that in a practical case, the EF is very nearly additive. For completely independent factors, the net mutual

TABLE 1. Relative Risk (RR), Attributable Fraction (AF), and Explained Fraction (EF) for Genetic Factors Affecting Slow Progression to AIDS

			AIDS Endpoint*											
			CD4 <200			AIDS 1993			AIDS 1987			Death		
	Model	Frequency	RR	AF	EF	RR	AF	EF	RR	AF	EF	RR	AF	EF
Protective factors														
CCR5-Δ32	dominant	20%	1.41	7.5%	1.1%	1.26	5.0%	0.5%	1.09	1.8%	0.1%	1.24	4.5%	0.7%
CCR2-64I	dominant	18%	1.24	4.1%	0.4%	1.11	1.9%	0.1%	1.06	1.1%	0.1%	1.14	2.5%	0.2%
SDF1-3A	recessive	4%	1.36	1.4%	0.2%	1.42	1.6%	0.3%	1.40	1.5%	0.6%	1.50	1.9%	0.9%
*HLA-B*27*	dominant	10%	1.28	2.8%	0.3%	1.47	4.5%	0.9%	1.51	4.9%	2.3%	1.41	4.0%	1.3%
*HLA-B*57*	dominant	9%	1.62	5.1%	1.3%	1.54	4.5%	1.0%	1.22	1.9%	0.4%	1.07	0.6%	<0.1%
KIR-3DS1—HLA-BW4 80I interaction	codominant	†	†	†	1.9%	†	†	2.3%	†	†	1.2%	†	†	1.7%
Susceptible factors														
CCR5-P1	recessive	12%	1.18	2.2%	0.4%	1.30	3.5%	1.1%	1.38	4.4%	0.8%	1.17	2.0%	0.2%
IL10-5A	dominant	44%	1.07	3.0%	0.1%	1.11	4.5%	0.3%	1.17	6.9%	0.4%	1.07	3.1%	0.1%
1× *HLA* Class I homozygosity	NA	9%	1.14	2.3%	0.3%	1.14	2.3%	0.3%	1.16	2.6%	0.2%	1.05	0.9%	<0.1%
2–3× *HLA* Class I homozygosity	NA	16%	1.30	1.4%	0.5%	1.25	1.2%	0.4%	1.47	2.2%	0.5%	1.48	2.3%	0.6%
*HLA-B*35Px*	dominant	5%	1.27	2.5%	0.7%	1.32	2.8%	1.0%	1.63	5.5%	1.7%	1.63	5.5%	1.9%
RANTES-H3/H5	dominant	6%	1.09	0.5%	0.1%	1.17	0.9%	0.2%	1.32	1.8%	0.3%	1.20	1.1%	0.1%
KIR-3DS1	codominant	†	†	†	0.9%	†	†	1.1%	†	†	0.8%	†	†	1.5%
All factors combined	additive	†	†	†	7.9%	†	†	8.7%	†	†	9.0%	†	†	8.3%
Sum of factor EFs					8.2%			9.3%			9.4%			9.3%

Subjects are 596 seroincident European Americans in 4 AIDS cohorts (Multicenter AIDS Cohort Study, San Francisco City Clinic Cohort , Multicenter Hemophilia Cohort Study, and AIDS Linked to the Intravenous Experience, as described.[6,18,19] To avoid confounding effects of Highly Active Antiretroviral Therapy, all outcomes are censored on 6/1/97 (ALIVE)) or 12/31/95 (other cohorts). RR and AF refer to slow progression for protective factors and to rapid progression for susceptible factors. Cutoff times for the 4 outcomes are as follows, respectively: for slow progression 9.0, 9.8, 10.7, and 12.1 years; for rapid progression 5.6, 5.5, 8.0 and 9.5 years.

*AIDS endpoint definitions: CD4 <200, first drop of $CD4^+$ cell count below 200 cell/μL; AIDS 1993, the CDC 1993 AIDS definition: clinical AIDS symptoms or CD4 <200; AIDS 1987, the CDC 1987 AIDS definition: clinical AIDS symptoms; death, AIDS-related death.

†Not calculated; frequencies and AF not well defined for multilevel factors.

NA indicates not applicable.

information is greater than the sum of the mutual information for individual factors due to the convexity of the information measure.[10] However, here some of the genetic factors are nonindependent, which tends to make mutual information smaller. The values for the AF calculated for individual factors are in every case much larger than the corresponding EF. We earlier calculated AFs for combined susceptible and combined protective factors (as noted above, there is no meaningful way to calculate an AF for both protective and combined factors); for slow progression to AIDS 1987, these were respectively 26.7% and 20.4%, both considerably larger than the EF for all factors.[19]

Table 2 gives the results for the corresponding calculation for rapid progression, defined as progression among the most rapid third of subjects. The overall EFs for rapid progression are somewhat smaller than for slow progression, ranging from 5% for progression to CD4 <200 to 7% for progression to AIDS 1987. Here in 2 cases, (CD4 <200 and AIDS 1993), the overall EF is larger than the sum of the individual factor EFs.

For comparison, we have calculated the EV for all 13 genetic factors considered together for the Cox model by the method of Schemper and Henderson.[8] Here we are measuring the influence of the genetic factors on overall progression considered as a continuous variable; hence, there is no distinction between early and rapid progression. The EV for 13 genetic factors, for progression to AIDS 1987, is 12.2%. The fact that the EV for genetic factors is somewhat larger than the EF may reflect the fact that the Cox model analysis captures early and late effects in the same analysis.

DISCUSSION

We have defined a statistic, the EF, which describes the extent to which cumulative categorical causal factors, in particular genetic factors, explain or predict categorical disease outcomes. We show that 13 ARGs have an EF of 9.0% for slow progression to AIDS. Thus, approximately 10% of differential progression is predicted or explained by these genetic factors. Although these fractions may seem small, it is notable that the EF for differential progression is similar to the EF (11.5%) for the influence of smoking on lung cancer.

From a public health standpoint, the AF is a natural measure of the impact of a causal factor on disease. Knowing that the AF of smoking for lung cancer death is 74%, we may assert that eliminating cigarette smoking would eliminate 74% of lung cancer deaths. However, the AF is not a good measure of our *knowledge* of the impact of cigarette smoking on lung cancer: the 74% figure obscures the fact that the great majority of smokers never get lung cancer, and thus incorrectly implies that the cause of lung cancer is largely understood. Moreover, the AF is not a particularly meaningful measure for the influence of genetic factors on disease, as these are not exposures that can be controlled. In addition, the

TABLE 2. Relative Risk (RR), Attributable Fraction (AF) and Explained Fraction (EF) for Genetic Factors Affecting Rapid Progression to AIDS

			AIDS Endpoint*											
			CD4 <200			AIDS 1993			AIDS 1987			Death		
	Model	Frequency	RR	AF	EF	RR	AF	EF	RR	AF	EF	RR	AF	EF
Protective factors														
CCR5-Δ32	dominant	20%	1.15	2.9%	0.6%	1.09	1.7%	0.3%	1.05	1.0%	0.1%	1.12	2.4%	0.5%
CCR2-64I	dominant	18%	1.09	1.6%	0.2%	1.04	0.7%	0.0%	1.03	0.6%	0.0%	1.07	1.3%	0.2%
SDF1-3A	recessive	4%	1.13	0.5%	0.1%	1.13	0.5%	0.2%	1.20	0.8%	0.5%	1.25	0.9%	0.6%
*HLA-B*27*	dominant	10%	1.11	1.1%	0.2%	1.14	1.4%	0.4%	1.25	2.4%	1.7%	1.21	2.0%	0.9%
*HLA-B*57*	dominant	9%	1.21	1.8%	0.7%	1.16	1.4%	0.5%	1.11	1.0%	0.3%	1.04	0.3%	0.0%
KIR-3DS1-HLA-BW4 80I interaction	codominant	†	†	†	1.0%	†	†	1.2%	†	†	0.9%	†	†	1.2%
Susceptible factors														
CCR5-P1	recessive	12%	1.27	3.2%	0.3%	1.49	5.6%	0.7%	1.47	5.4%	0.6%	1.21	2.4%	0.1%
IL10-5A	dominant	44%	1.10	4.1%	0.1%	1.16	6.4%	0.2%	1.20	8.0%	0.3%	1.09	3.6%	0.1%
1× *HLA* Class I homozygosity	NA	9%	1.21	3.3%	0.2%	1.21	3.4%	0.2%	1.19	3.1%	0.2%	1.06	1.0%	0.0%
2–3× *HLA* Class I homozygosity	NA	16%	1.48	2.3%	0.4%	1.41	2.0%	0.2%	1.59	2.8%	0.4%	1.62	2.9%	0.5%
*HLA-B*35Px*	dominant	5%	1.42	3.8%	0.5%	1.53	4.6%	0.7%	1.82	7.0%	1.4%	1.84	7.2%	1.6%
RANTES-H3/H5	dominant	6%	1.13	0.7%	0.0%	1.26	1.4%	0.1%	1.39	2.2%	0.2%	1.24	1.3%	0.1%
KIR-3DS1	codominant	†	†	†	0.5%	†	†	0.6%	†	†	0.6%	†	†	1.1%
All factors combined	additive	†	†	†	5.0%	†	†	5.5%	†	†	7.0%	†	†	6.5%
Sum of factor EFs					4.7%			5.2%			7.1%			7.0%

Subjects are 596 seroincident European Americans in 4 AIDS cohorts (Multicenter AIDS Cohort Study, San Francisco City Clinic Cohort , Multicenter Hemophilia Cohort Study, and AIDS Linked to the Intravenous Experience), as described.[6,18,19] To avoid confounding effects of Highly Active Antiretroviral Therapy, all outcomes are censored on 6/1/97 (ALIVE) or 12/31/95 (other cohorts). RR and AF refer to slow progression for protective factors and to rapid progression for susceptible factors. Cutoff times for the 4 outcomes are as follows, respectively: for slow progression 9.0, 9.8, 10.7, and 12.1 years; for rapid progression 5.6, 5.5, 8.0 and 9.5 years.

*AIDS endpoint definitions: CD4 <200, first drop of $CD4^+$ cell count below 200 cell/μL; AIDS 1993, the CDC 1993 AIDS definition: clinical AIDS symptoms or CD4 <200; AIDS 1987, the CDC 1987 AIDS definition: clinical AIDS symptoms; death, AIDS-related death.

†Not calculated; frequencies and AF not well defined for multilevel factors.

NA indicates not applicable.

definition of the AF requires dividing the subjects into exposed and unexposed, but for genetic factors, it is often arbitrary which allele of a locus constitutes the exposure, and in addition, this approach does not deal effectively with the case of multiple factors.

For scientific understanding, it is more useful to know to what extent the proven causal factors predict the outcomes. This is naturally measured by the correlation between factors and outcomes, measured by R^2 for linear regression or by the EV for more general regression. The EF is the natural extension of the EV to a contingency table, and thus provides a statistic that applicable to the cases for which an AF is defined. By making simplifying assumptions, both confidence intervals and good approximations to true population value of the EF can be calculated (Appendix 2).

As the AF has the practical application of guiding public health measures to reduce exposures, the EF or the EV, as measures of explanation, have a potentially important application to allocating research resources by giving an approximate measure of how much of the causality of a disease is unknown. Unknown factors include environmental, host-genetic, pathogen-genetic, and stochastic effects. For the case of AIDS progression, as for the case of smoking, the AF is much larger than the EF, so the AF, taken as a measure of the degree of scientific understanding, greatly overestimates how well the disease is understood. Although it is unknown how much of AIDS progression is determined by genetic factors, it is plausible that it is substantially more than 9% EF that we have calculated for the known ARGs, suggesting that there is significant genetic influence on this disease that remains to be discovered.

ACKNOWLEDGMENTS

We would like to thank Sahu Ashutosh for technical assistance and Mitch Gail, Raymond Carroll, and Karen Kessler for useful discussions.

REFERENCES

1. Benichou J. Attributable Risk. In: Mitchell H, Gail JB eds. *Encyclopedia of Epidemiologic Methods. Vol 1*. NY: Wiley, 2000:216–229.
2. Levin ML. The occurrence of lung cancer in man. *Acta Union Int Contra Cancrum*. 1953;9:531–541.
3. Dean M, Carrington M, Winkler C, et al. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study. *Science*. 1996;273(5283):1856–1862.
4. Blanpain C, Libert F, Vassart G, et al. CCR5 and HIV infection. *Receptors Channels*. 2002;8(1):19–31.
5. Bruzzi P, Green SB, Byar DP, et al. Estimating the population attributable risk for multiple risk factors using case-control data. *Am J Epidemiol*. 1985;122(5):904–914.
6. Silverberg MJ, Smith MW, Chmiel JS, et al. Fraction of cases of acquired immunodeficiency syndrome prevented by the interactions of identified restriction gene variants. *Am J Epidemiol*. 2004;159(3):232–241.
7. Schemper M. Predictive accuracy and explained variation. *Stat Med*. 2003;22(14):2299–2308.

8. Schemper M, Henderson R. Predictive accuracy and explained variation in Cox regression. *Biometrics*. 2000;56(1):249–255.
9. Schemper M, Stare J. Explained variation in survival analysis. *Stat Med*. 1996;15(19):1999–2012.
10. Kullback S. *Information Theory and Statistics*. Mineola, NY: Dover, 1968.
11. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J*. 1948;27:379–423. 623–656.
12. Fredrickson DS, Altrocchi PH, Avioli OV, et al. Tangier disease—combined clinical staff conference at the National Institutes of Health. *Ann Intern Med*. 1961;55:1016–1031.
13. Rust S, Walter M, Funke H, et al. Assignment of Tangier disease to chromosome 9q31 by a graphical linkage exclusion strategy. *Nat Genet*. 1998;20(1):96–98.
14. Langmann T, Klucken J, Reil M, et al. Molecular cloning of the human ATP-binding cassette transporter 1 (hABC1): evidence for sterol-dependent regulation in macrophages. *Biochem Biophys Res Commun*. 1999;257(1):29–33.
15. Bodzioch M, Orso E, Klucken J, et al. The gene encoding ATP-binding cassette transporter 1 is mutated in Tangier disease. *Nat Genet*. 1999;22(4):347–351.
16. Nilsson S, Carstensen JM, Pershagen G. Mortality among male and female smokers in Sweden: a 33 year follow up. *J Epidemiol Community Health*. 2001;55(11):825–830.
17. Proctor RN. Tobacco and the global lung cancer epidemic. *Nat Rev Cancer*. 2001;1(1):82–86.
18. O'Brien SJ, Nelson GW. Human genes that limit AIDS. *Nat Genet*. 2004;36(6):565–574.
19. O'Brien SJ, Nelson GW, Winkler CA, et al. Polygenic and multifactorial disease gene association in man: lessons from AIDS. *Annu Rev Genet*. 2000;34:563–591.
20. Carrington M, O'Brien SJ. The influence of HLA genotype on AIDS. *Annu Rev Med*. 2003;54:535–551.
21. Dean M, Carrington M, O'Brien SJ. Balanced polymorphism selected by genetic versus infectious human disease. *Annu Rev Genomics Hum Genet*. 2002;3:263–292.
22. Liao JG, Mcgee D. Adjusted coefficients of determination for logistic regression. *Am Stat*. 2003;57(3):161–165.
23. Sokal RR, Rohlf FJ. *Biometry*, 3rd ed. New York: Freeman and Company, 1995.
24. Mittlbock M, Schemper M. Computing measures of explained variation for logistic regression models. *Comput Methods Programs Biomed*. 1999;58(1):17–24.

APPENDIX 1: ANALOGY WITH R^2 AND EV

We may write the definition (4) of the EF as:

$$EF = \frac{\sum_{ij} a_{ij}\left(\log\left(\frac{a_{ij}}{a_{i\cdot}}\right) + \log\left(\frac{1}{a_{\cdot j}}\right)\right)}{-\sum_j a_{\cdot j}\log(a_{\cdot j})} = \frac{-\sum_j a_{\cdot j}\log(a_{\cdot j}) + \sum_i a_{i\cdot}\sum_j \frac{a_{ij}}{a_{i\cdot}}\log\left(\frac{a_{ij}}{a_{i\cdot}}\right)}{-\sum_j a_{\cdot j}\log(a_{\cdot j})}. \quad \text{(A1)}$$

Thus,

$$EF = 1 - \frac{-\sum_i a_{i\cdot}\sum_j \mathrm{p}_{ij}\log(\mathrm{p}_{ij})}{-\sum_j a_j\log(a_j)}. \quad \text{(A2)}$$

Here p_{ij} is the probability that an individual carrying factors i falls into disease category j; thus $-\sum_j pij\log(pij)$ is the entropy of row i, and the sum over the rows weighted by the frequencies a_i of the factors gives the residual entropy—the residual uncertainty of disease state—with knowledge of the disease factors. It may be seen that the EF is analogous to the regression R^2,[22] and thus is in effect an EV for the contingency table. However, we have used the definition (4) and the nomenclature "explained fraction" (EF), as these emphasize the analogy with the attributable factor (AF). Logically, the name "explained fraction" should be reserved for calculations based on population frequencies, that is, cohort studies rather than case-control studies. Indeed, any explained variation for a population could be considered to be an explained fraction.

APPENDIX 2: STATISTICS OF THE EXPLAINED FRACTION

Although the EF is closely related to the likelihood ratio statistic, our use of it is quite different, leading to different statistical questions. The likelihood ratio statistic is used to determine whether the rows and columns of a matrix are significantly nonindependent, using the fact that under the assumption of independence, 2 times the statistic is asymptotically distributed as a χ^2.[23] For us, the object of interest is the true value of $I(1:2)$ for the underlying population; thus, we need estimators of this parameter from the observed values. The observed value $I(1,2)$, that is, the value calculated by Eq. (4) using table frequencies obtained from a particular sample, is a biased estimator for $I(1,2)$. This bias is clearly shown by two special cases. First, if there is no correlation between factor and disease in the population, the true value of $I(1:2)$ is 0. However, most matrices of samples of this population will have nonzero $I(1:2)$; because this is always positive, the mean value of the observed $I(1,2)$ will be greater than 0. Second, and more critical, consider the case where we have a very large set of genetic markers. With enough (unlinked) markers, there will only be a single individual for each composite genotype (or no individuals, in which case, the row may be ignored). In this case, the genotype necessarily and trivially "predicts" the disease group for each individual, and EF = 1, even if the genetic markers have no causal relation with the disease at all; that is, the model is saturated. Practically, we must take different approaches to the case of many factors; the calculation above of the effect of genetic factors on AIDS progression shows one approach. In this section, we consider the estimation of the EF and its confidence interval.

The null hypothesis of independence of the rows and columns of a matrix is equivalent to the assumption that $I(1:2) = 0$ in the population. For $I(1:2) \neq 0$ in the population, for N subjects observed, $2N$ times the observed $I(1,2)$ is asymptotically distributed as a noncentral χ^2 with noncentrality parameter $2NI(1,2)$ and $(n-1)(m-1)$ degrees of freedom for n rows and m columns.[10] Thus, we have a likelihood function,

$$L(2NI(1:2)) = \chi^2_{2NI(1:2)}(2N\hat{I}(1:2)), \quad \text{(B1)}$$

from which we can obtain a maximum likelihood estimation and confidence intervals. Frequently, the contingency

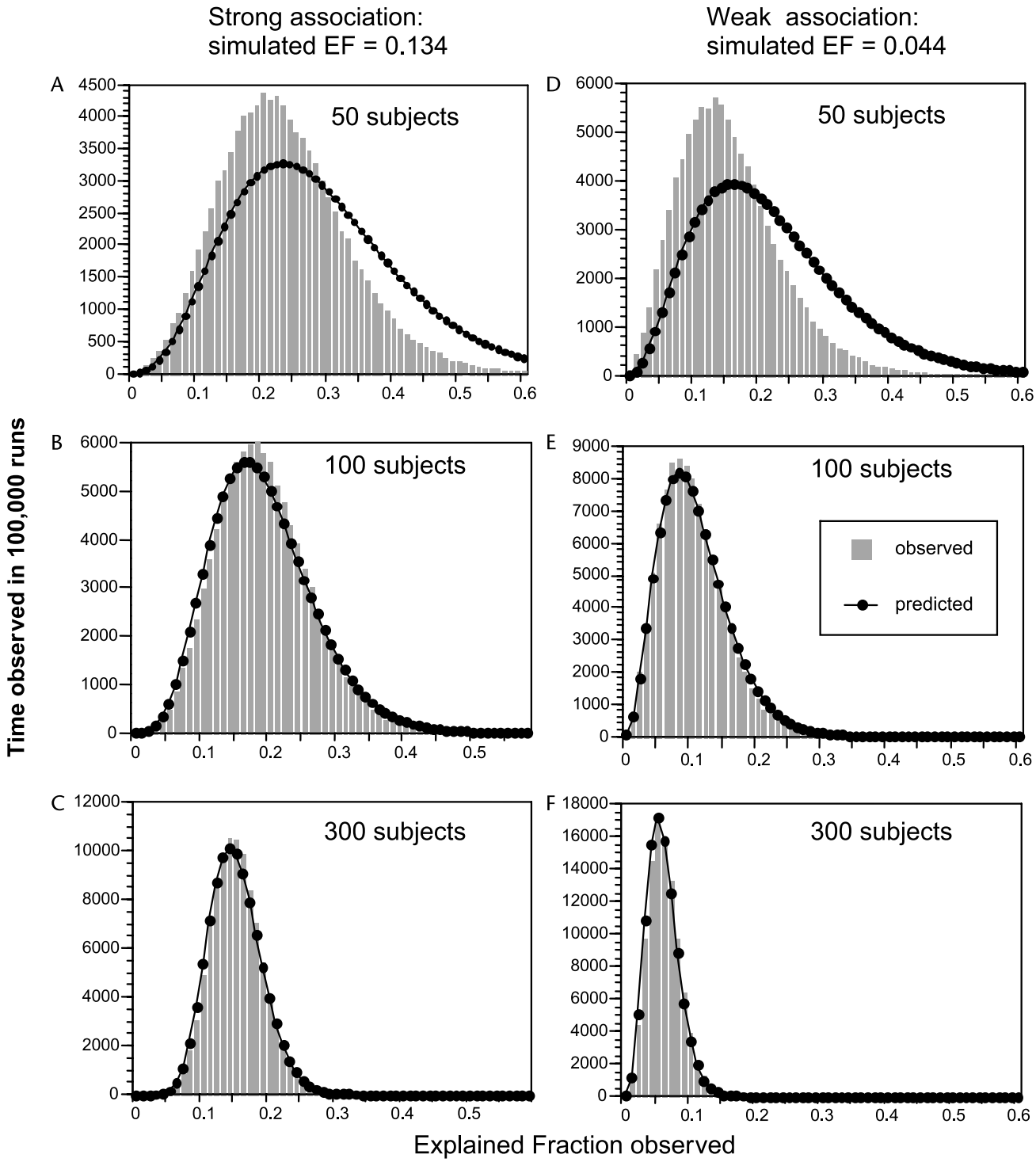

FIGURE 1. Simulation results: Histograms of explained fraction (EF) calculated from contingency tables drawn randomly from assumed population frequencies (bars), compared with noncentral χ^2 prediction of observed EF (circles). Vertical scale is number of subjects in a 0.01 interval of observed EF. Assumed population contingency tables are for scenarios with true EF of 13.4% (A, B, C), or 4.4% (D, E, F), as described in the text. Random samples are of 50 (A, D), 100 (B, E), or 300 (C, F) individuals. A total of 100,000 random samplings were performed for each case. Note change of vertical scale between plots.

TABLE 3. Comparison of Mean Observed Explained Fraction (EF) With Maximum Likelihood (ML) Estimate of EF From the Noncentral χ^2 Likelihood Function, for 4 Simulated Sample Sizes From Assumed Population Frequencies for Strong and Weak EF Scenarios

Simulated EF	N	Mean EF Observed	Mean ML Estimated EF	% Error In Observed	% Error In Estimated	Test of 95% CI	
						% Below	% Above
13.4%	50	23.6%	13.2%	77%	−1.5%	0.7%	2.4%
13.4%	100	19.3%	14.2%	44%	5.6%	1.5%	2.2%
13.4%	300	15.5%	13.8%	16%	3.5%	2.1%	1.7%
13.4%	1000	14.0%	13.5%	4.5%	0.9%	1.9%	1.9%
4.4%	50	15.6%	5.7%	258%	31%	1.2%	<0.1%
4.4%	100	10.5%	5.4%	141%	24%	2.4%	0.5%
4.4%	300	6.5%	4.8%	48%	9.5%	2.8%	2.2%
4.4%	1000	5.0%	4.5%	14%	2.4%	2.6%	2.4%

CI indicates confidence interval.

table will be sparse, and χ^2 may not be a good approximation. Also, to apply Eq. (B1) to the EF, we must make the additional approximation that the EF varies proportionally to the mutual information, although in fact, the denominator of the EF is also varying. However, we show in the simulations below that for a realistic case, these approximations allow a reasonably accurate estimate of the true population EF.

Limiting the Degrees of Freedom

When many factors influence the outcome, the number of combinations of factors becomes very large, and as noted, direct computation of the EF from Eq. (7) becomes meaningless. Here an alternative is to use an explained variation measure calculated by logistic regression.[24] To obtain a meaningful EF measure for the contingency table, we may make a simplifying approximation that the effects of the different factors are independent, so that the relative risks of combinations of factors are the products of the relative risks of the individual factors. For cases where factors have significant interactions, we add new effects representing the interaction, so that an enlarged set of independently acting factors is produced.

Formally, if we have M factors, we can consider the contingency table in its alternate form of an $M + 1$ dimensional table (one dimension for each factor, and one for disease state). With the additional assumption that the factors occur independently in the population, the assumption of independence of effects implies that in the underlying population, for a case of 3 explanatory factors indexed by i, p, and q, $\alpha_{ipqj} = \alpha_{i\cdot\cdot j}\alpha_{\cdot p\cdot j}\alpha_{\cdot\cdot qj}$, where the α's are the entries in the multidimensional frequency table, or the corresponding marginals. Therefore, to estimate the EF by this method, we replace the observed $\hat{\alpha}_{ipqj}$ by $\hat{\alpha}_{i\cdot\cdot j}\hat{\alpha}_{\cdot p\cdot j}\hat{\alpha}_{\cdot\cdot j}$. In particular circumstances, better approximations to the effects of combined factors may be available; we use one such approach below for the case of genetic factors affecting progression to AIDS. The critical issue is to reduce the degrees of freedom.

Simulation: Test of Bias in the Observed EF

To illustrate the bias and variance of the EF that would be calculated from observations of actual population frequencies, we simulate observations of a specific disease association scenario. We suppose that there are 2 disease categories, D_1 and D_2, with frequencies 0.7 and 0.3, and 3 causal factors with population frequencies 0.4, 0.2, and 0.1, that are uncorrelated in the population and act independently, so that the relative risk for multiple factors is the product of the risk for individual factors. We first suppose that the relative risks for the 3 factors are 2.0, 0.5, and 3.0, respectively. By Eq. (4), the EF for the 3 factors combined is 13.4%. We also simulate a case of weaker factors, with the 3 factors having the same frequencies but relative risks now of 1.5, 0.7, and 2.0, respectively, yielding an EF of 4.4%.

We simulate studies that observe either 50, 100, 300, or 1000 subjects from this population. For each case, we run 100,000 random draws, and for each draw, we calculate an observed EF, a maximum likelihood estimate of the EF based on the noncentral χ^2 distribution with 7 degrees of freedom, for the observed frequencies, and upper and lower 95% confidence limits for this distribution. Figure 1A-F shows histograms comparing the simulation results for 50, 100, or 300 subjects to the predicted noncentral χ^2 distribution of the EF. The noncentral χ^2 gives an excellent fit for 100 or more subjects. Table 3 gives, for the same simulations, the mean of the observed EF, the mean of the maximum likelihood estimates of the EF, and the frequencies of the actual EF being above or below the calculated confidence interval. We note that the maximum likelihood (ML) estimate of the EF is much less biased than the observed EF; for most cases, the mean of the maximum likelihood estimate is quite close to the actual EF, and the frequency with which the true value is above or below the 95% confidence intervals is close to 2.5%.